### Authors:

Article title:<br>Estimated ge<br>Short title: E<br>Authors:<br>Dorothée D<br>Crawford<sup>4</sup>, T<br>1 Genetics ar<br>3 Genetics ar<br>4 Global Evid Estimated genetics prevalence of *PRKN-PD*<br>Short title: Estimated genetics prevalence of *PRKN-PD*<br>Authors:<br>Dorothée Diogo<sup>1</sup>, Emily H.M. Wong<sup>2,8</sup>, Christopher DeBoever<sup>2</sup>, Wei Qu<sup>3</sup>, Jonghun Lec<br>Crawford<sup>4</sup>, Takeshi Hiok ミー・ノー I C フィー・ニュー E E E Short title: Estimated genetics prevalence of PARN-1 D<br>
Dorothée Diogo<sup>1</sup>, Emily H.M. Wong<sup>2,8</sup>, Christopher<br>
Crawford<sup>4</sup>, Takeshi Hioki<sup>5</sup>, Jaya Padmanabhan<sup>6</sup>, Daria P<br>
1 Genetics and Systems Biology, Takeda Development<br> 

Dorothée Diogo<sup>1</sup><br>Crawford<sup>4</sup>, Takesł<br>1 Genetics and Sy<br>2 Genetics and Sy<br>3 Genetics and Sy<br>4 Global Evidence<br>5 Neuroscience D<br>6 Neuroscience D<br>02142<br>7 Neuroscience D<br>8 Current address , Emily H.M. Wong<sup>2,3</sup>, Christopher DeBoever<sup>2</sup><br>11 Hioki<sup>5</sup>, Jaya Padmanabhan<sup>6</sup>, Daria Prilutsky<sup>1</sup>, Ga<br>12 stems Biology, Takeda Development Center Amer<br>12 stems Biology, Takeda Development Center Amer<br>12 stems Biology, , Wei Qu<sup>3</sup><br>briele Proet<br>icas, Inc., C<br>icas, Inc., S<br>imited, Fuj<br>ericas, Inc.,<br>any Limited<br>enter Amer<br>Americas, Inc., San Fr , Jonghun Lee<sup>1</sup><br>1722 Jambridge, MA<br>1822<br>ambridge, MA<br>1832 Wa, Japan<br>Cambridge, MA<br>1635, Inc., Camb<br>1628, Inc., Camb<br>1628, Inc., Camb<br>1628, Napar<br>1628, Napar 2139<br>121<br>121<br>n.<br>ridge, MA<br>14 02142<br>38 Crawford\*<br>1 Genetics<br>2 Genetics<br>3 Genetics<br>4 Global E<br>5 Neurosc<br>02142<br>7 Neurosc<br>8 Current , Takeshi Hioki<sup>-</sup><br>Fand Systems B<br>and Systems B<br>and Systems B<br>vidence and Ou<br>ience Drug Disc<br>cience Therape<br>address: Data S<br>address: Data S , Jaya Padmanabhan<br>iology, Takeda Devek<br>iology, Takeda Devek<br>iology, Takeda Pharm<br>icomes, Takeda Deve<br>covery Unit, Takeda Pl<br>utic Area Unit, Take<br>covery Unit, Takeda Deve , Daria Prilutsky<br>Dopment Center A<br>Dopment Center A<br>Raceutical Compa<br>Popment Center<br>Rarmaceutical Coda<br>Development Center<br>Revelopment Center<br>S. Vir Biotechnol , Gabriele Proetzel´<br>
.mericas, Inc., Camk<br>
.mericas, Inc., San D<br>
.my Limited, Fujisaw<br>
Americas, Inc., Cam<br>
.mpany Limited, Fuj<br>
t Center Americas,<br>
ter Americas, Inc., C<br>
ogy Inc., San Franci 2 Genetics and Systems Biology, Takeda Development Center Americas, Inc., San Diego, CA 92121<br>3 Genetics and Systems Biology, Takeda Pharmaceutical Company Limited, Fujisawa, Japan.<br>4 Global Evidence and Outcomes, Takeda D 2 Genetics and Systems Biology, Takeda Pharmaceutical Company Limited, Fujisawa, Japan.<br>2 Genetics and Systems Biology, Takeda Pharmaceutical Company Limited, Fujisawa, Japan.<br>4 Global Evidence and Outcomes, Takeda Develop 3 General General Company Commission Control Company Limited, Fujisawa, Japan. Sheuroscience Drug Discovery Unit, Takeda Pharmaceutical Company Limite 4 Global Evidence Drug Discovery Unit, Takeda Pharmaceutical Company Limited, Fujisawa, Japan.<br>
5 Neuroscience Therapeutic Area Unit, Takeda Development Center Americas, Inc., Cambridge,<br>
02142<br>
7 Neuroscience Drug Discove 5 Neuroscience Therapeutic Area Unit, Takeda Development Center Americas, Inc., Cambrid<br>02142<br>7 Neuroscience Drug Discovery Unit, Takeda Development Center Americas, Inc., Cambridge, MA<br>8 Current address: Data Science and

02142<br>
6 Neuroscience Drug Discovery Unit, Takeda Development Center Americas, Inc., Cambridge, MA 02142<br>
8 Current address: Data Science and Operations, Vir Biotechnology Inc., San Francisco, CA 94158<br> **Corresponding auth** 

### Corresponding author:

7 Neur<br>8 Curre<br>Corres<br>Doroth<br>35 Lan:<br>doroth 8 Current address: Data Science and Operations, Vir Biotechnology Inc., San Francisco, CA 94158<br>Corresponding author:<br>Dorothée Diogo<br>35 Lansdowne street, Cambridge, MA 02139, USA<br>dorothee.diogo@takeda.com Expression and Current and Current and Operations, Vir Biotechnology Inc., Corresponding author:<br>Dorothée Diogo<br>35 Lansdowne street, Cambridge, MA 02139, USA<br>dorothee.diogo@takeda.com  $\frac{1}{2}$ Bunsdowne st<br>dorothee.diogo@<br>Dorothee.diogo@ 35 Lansdowne street, Cambridge, MA 12213<br>dorothee.diogo@takeda.com

### **Abstract**

Abstract<br>Backgrou<br>Estimatin<br>of studie:<br>estimate<br>variants i<br>Methods<br>We cura<br>frequenci<br>used the Estimating the<br>of studies in<br>estimate the<br>variants in the<br>Methods<br>We curated<br>frequencies<br>used the car<br>Results of studies in diverse populations, and frequent under or misdiagnosis. We leveraged human genetics to<br>estimate the genetic prevalence (eGP) of familial Parkinson's disease (PD) caused by biallelic pathogenic<br>variants in th

of studies the genetic previalence (eGP) of familial Parkinson's disease (PD) caused by biallelic pathogenic<br>variants in the *Parkin (PRKN)* gene (PRKN-PD).<br>Methods<br>We curated the reported *PRKN*-PD pathogenic variants and

wariants in the *Parkin* (*PRKN*) gene (*PRKN*-PD).<br>
Methods<br>
We curated the reported *PRKN-PD* pathogenic variants and obtained the heterozygous carrier<br>
frequencies of these variants from gnomAD and the Japanese Multi-o Methods<br>We curated the reported *PRKN*-PD pathog<br>frequencies of these variants from gnomAD an<br>used the carrier frequencies to estimate the e(<br>Results<br>Non-Japanese East Asians presented the highe<br>165 per 100,000 individuals 『\ f u \_ F r c u We cura<br>frequenci<br>used the of<br>Results<br>Non-Japa<br>165 per 1<br>64 per 10<br>the USA a<br>and 13 p<br>were sig The random the reported Finance is pathogenic current and contract the interterported process of these variants from gnomAD and the Japanese Multi-omics reference panel (jMorp). We used the carrier frequencies to estimate frequencies of these variants from graduate the eGP of *PRKN-PD* in eight genetic ancestries.<br>Results<br>Non-Japanese East Asians presented the highest eGP of *PRKN-PD* (24 per 100,000 individuals, 95% Cl=4-<br>165 per 100,000 Results<br>
Non-Japanese East Asians presented the highest eGP of *PRKN*-PD (24 per 100,000 individuals)<br>
165 per 100,000 individuals), followed by Non-Finnish Europeans (22 in 100,000 individuals)<br>
64 per 100,000 individuals Non-Jap<br>165 per<br>64 per 1<br>the USA<br>and 13<br>were sindividu<br>Conclus<br>This is<br>prevalen The First Asian Markov and Markov and Markov and States (22 in 100,000 individuals, 95% CI = 11-64 per 100,000 individuals), followed by Non-Finnish Europeans (22 in 100,000 individuals, 95% CI = 11-64 per 100,000 individu 64 per 100,000 individuals). Based on the proportions of races and ethnicities, we estimated the eGP in<br>the USA and the world-wide eGP to be 18 per 100,000 individuals (95% CI=7-68 per 100,000 individuals).<br>and 13 per 100,

the USA and the world-wide eGP to be 18 per 100,000 individuals (95% Cl=7-68 per 100,000 individuals).<br>and 13 per 100,000 individuals (95% Cl=3-70 per 100,000 individuals), respectively. These estimates<br>were significantly and 13 per 100,000 individuals (95% CI=3-70 per 100,000 individuals), respectively. These estimates were significantly reduced when excluding structural variants (world-wide eGP=2 per 100,000 individuals, 95% CI=1-5 per 10 and 13 per 100,000 individuals (13 per 100,000 individuals).<br>Were significantly reduced when excluding structural variants (world-wide eGP=2 per 100,000<br>individuals, 95% CI=1-5 per 100,000 individuals).<br>Conclusions<br>This is individuals, 95% Cl=1-5 per 100,000 individuals).<br>
Conclusions<br>
This is the first study estimating the *PRKN-PD* genetic prevalence. Our results suggest that the<br>
prevalence of the disease may be higher than previously re Conclusions<br>This is the first study estimating the *PRKN*-F<br>prevalence of the disease may be higher than pre<br>We also demonstrate the importance of careful<br>disease, and its limitations, when using the ap<br>genetic prevalence. ( ヿ ド ヽ g This is the<br>prevalence c<br>We also den<br>disease, and<br>genetic prev The first study of the disease may be higher than previously reported, highlighting potential underdiagnosis.<br>We also demonstrate the importance of carefully considering the known genetic epidemiology of each<br>disease, and previous of the disease may be higher many be higher may be previously reported, highlighting potential underdiagnosis.<br>We also demonstrate the importance of carefully considering the known genetic epidemiology of each<br>dis We also demonstrate the importance of carefully considered in this study to estimate the disease genetic prevalence.<br>disease, and its limitations, when using the approach applied in this study to estimate the disease<br>genet disease, and its limitations, when using the applied in this study to estimate the disease of  $\frac{1}{2}$ genetic prevalence.

ノミ N H E S n c c As of 2022, about 7000 rare diseases have been identified<br>30 million people in the United States (US) and more tha<br>rare diseases have FDA-approved drug treatments, highl<br>patients living with these diseases<sup>2</sup>.<br>Robust epide n 300 million people worldwide. Only 5% of<br>ighting the high unmet need remaining for<br>ssess the burden of a rare disease on patients<br>crucial to help guide clinical development<br>a disease is in a population by tallying the<br>Ho Example in the Entitled States (1, paid that is mailed states finited states (discussion people in the United States (on patients living with these diseases<sup>2</sup>.<br>Robust epidemiology data is critically important to better as rare interesting with these diseases<sup>2</sup>.<br>Robust epidemiology data is critically important to better assess the burden of a rare disease on patients<br>and families<sup>3</sup>. In drug development, this information is crucial to help patients living with these diseases<sup>2</sup><br>Robust epidemiology data is critica<br>and families<sup>3</sup>. In drug developme<br>strategies. Traditionally, one can<br>number of diagnosed patients ir<br>diagnosed, leading to unreliable<br>diseases, wh . Illy an m n d its a c p f in ト ミ ミ r く く ド ノ ド ト and families<sup>3</sup>. In drug development, this information is crucial to help guide clinical development<br>strategies. Traditionally, one can measure how common a disease is in a population by tallying the<br>number of diagnosed pa and families<br>strategies. Tr<br>number of d<br>diagnosed, le<br>diseases, whe<br>disease, the f<br>valuable met<br>prevalence, i.<br>Assuming full<br>probability of<br>have used th<br>example, Pon The image of the population is a disease is in a population by tallying the liagnosed patients in the population itself. However, many rare diseases are under-<br>aditionally, one can measure how common a disease. For rare au strategies. Traditionally, the transmission itself. However, many rare diseases are under-<br>diagnosed, leading to unreliable disease prevalence estimates. For rare autosomal recessive (AR)<br>diseases, where pathogenic variant diseases, where pathogenic variants in both alleles (maternal and paternal) are required to cause the disease, the frequency of known pathogenic variants in large reference population datasets serve as a valuable metric to meters, the frequency of known pathogenic variants in large reference population datasets serve as a<br>disease, the frequency of known pathogenic variants in large reference population datasets serve as a<br>valuable metric to metric is estimate how common a disease may be. This approach estimates the genetic<br>prevalence, i.e., the proportion of people with specific causal pathogenic genotypes in the population.<br>Assuming full disease penetrance o **Example metric in a disease. Several published studies have used this appr** 

Familial Parkinson's disease (PD) is an autosomal recessive (AR) condition caused by pathogenic variants previously and preparation people and people with people general general pathogen as<br>Assuming full disease penetrance of the variants, the genetic prevalence refers to the lifetime risk – the<br>probability of an individual ( **Example 18 The variant full disease.** Several published studies<br>
have used this approach to predict the genetic prevalence of rare recessive diseases, including, for<br>
example, Pompe disease, Wilson disease, inherited reti make used this approach to predict the genetic prevalence of rare recessive diseases, including, for example, Pompe disease, Wilson disease, inherited retinal diseases, ENPP1 deficiency and congenital thrombotic thrombocyt example, Pompe disease, Wilson disease, inherited retinal diseases, ENPP1 deficiency and congenital<br>thrombotic thrombocytopenic purpura<sup>4-8</sup>.<br>Familial Parkinson's disease (PD) is an autosomal recessive (AR) condition cause example, Pompe disease, Wilson disease, inherited retinal diseases, ENPP1<br>thrombotic thrombocytopenic purpura<sup>4-8</sup>.<br>Familial Parkinson's disease (PD) is an autosomal recessive (AR) condition caus<br>in the *Parkin* (*PRKN*) g deficiency<br>ed by path<br>ins 12 exor<br>ing substra<br>targeting<br>depolarized<br>essive fam<br>a at onset<br>a, resting i<br>cally, it is d and congenital<br>ogenic variants<br>s and is one of<br>te proteins for<br>for regulating<br>l mitochondria<br>lial Parkinson's<br>being 31 years<br>remor, rigidity,<br>fferentiated by thrombotic<br>Familial Par<br>in the *Parki*<br>the largest<br>proteasoma<br>mitochondr<br>(mitophagy<br>Disease<sup>9</sup>. Cl<br>old), slowly<br>and clinicall thrombocytopenic<br>kinson's disease (P<br>n (PRKN) gene<sup>9</sup>. PR<br>human genes. It e<br>il, and is one of<br>ial quality control<br>j<sup>10</sup>. Biallelic loss-of<br>inically, PRKN-PD l<br>progressive and cl<br>y and pathologicall purpura<sup>4</sup><br>
D) is an auto<br>
XN is locate<br>
Photodes an<br>
the two<br>
and prom<br>
function va<br>
manifests w<br>
assical moto<br>
y distinct fr FitFr(Ccc in the *Parkin* (*PRKN*) gene<sup>8</sup>. *PRKN* is located on chromosome 6q25.2-27, contains 12 exons and is one of<br>the largest human genes. It encodes an E3 ubiquitin ligase critical in targeting substrate proteins for<br>proteaso in the *Parkin* (*PRKN*) gene<sup>3</sup><br>the largest human genes.<br>proteasomal, and is one<br>mitochondrial quality con<br>(mitophagy)<sup>10</sup>. Biallelic loss<br>Disease<sup>9</sup>. Clinically, *PRKN*-lold), slowly progressive an<br>and clinically and pat The three contents of the three contents in the three contents of the three codes an E3 ubiquitin ligase critical in targeting substrate proteins for of the two kinases that has a mitochondrial targeting for regulating tro proteasomal, and is one of the two kinases that has a mitochondrial targeting for regulating<br>mitochondrial quality control and promoting the selective autophagy of depolarized mitochondria<br>(mitophagy)<sup>10</sup>. Biallelic lossmitochondrial quality control and promoting the selective autophagy of depolarized mitochondria (mitophagy)<sup>10</sup>. Biallelic loss-of function variants in *PRKN* cause autosomal recessive familial Parkinson's Disease<sup>9</sup>. Cli (mitophagy)<sup>10</sup>. Biallelic loss-of function variants in *PRKN* cause autosomal recessive familial Parkinson's<br>Disease<sup>9</sup>. Clinically, *PRKN*-PD manifests with early disease onset (median age at onset being 31 years<br>old), (mitophagy)<sup>10</sup>. Biallelic loss-of function variants in *PRKN* cause autosomal recessive familial Parkinson's<br>Disease<sup>9</sup>. Clinically, *PRKN*-PD manifests with early disease onset (median age at onset being 31 years<br>old), s Disease<br>old), slov<br>and clini of the clinical progressive and classical motoric symptoms, including bradykinesia, resting tremor, rigidity,<br>cally and pathologically distinct from idiopathic PD<sup>11</sup>. Neuropathologically, it is differentiated by<br>3 and clinically and pathologically distinct from idiopathic  $PD^{11}$ . Neuropathologically, it is differentiated by 3 and clinically and pathologically distinct from idiopathic PD<sup>11</sup>. Neuropathologically, it is differentiated by<br>3<br>3

general above of the substantial angra and locus coeruleus<sup>12</sup>.<br> *PRKN*-PD has been suggested to represent 0.5-1% of individuals with PD across all age-at-onset (AAO)<br>
groups based on the proportions of *PRKN* pathogenic loss confined to the substantia nigra and locus coeruleus<sup>11</sup>.<br> *PRKN*-PD has been suggested to represent 0.5-1% of indiv<br>
groups based on the proportions of *PRKN* pathogenic<br>
studies<sup>13</sup>. Together with the estimated tota groups based on the proportions of *PRKN* pathogenic variants carriers observed in published PD<br>studies<sup>13</sup>. Together with the estimated total number of PD patients in the world<sup>34</sup>, the number of *PRKN*-<br>PD patients is e groups and the proportions of the estimated on the proportions in the world<sup>14</sup>, the number of PRKN-<br>PD patients is estimated at 35,000-70,000 worldwide<sup>13</sup>. PRKN pathogenic variants are more prevalent<br>among adolescent PD studies". Together with the estimated total number of PD patients in the world", the number of PRKN-<br>PD patients is estimated at 35,000-70,000 worldwide<sup>13</sup>. PRKN pathogenic variants are more prevalent<br>among adolescent PD PD patients is estimated at 35,000-70,000 worldwide<sup>22</sup>. PRKN pathogenic variants are more prevalent<br>among adolescent PD patients (AAO<20 years old, 77%) and adult patients with AAO<=50 years old (2.6-<br>2.8%)<sup>35-17</sup>. Howeve 2.8%)<sup>15-17</sup>. However, prevalence estimates from these studies might be affected by ascertainment bias, including different proportions of familial and idiopathic patients recruited, and small sample sizes. In addition, an

2.8%)<sup>916</sup>. However, prevalence estimates from these studies might be affected by ascertainment bias, including different proportions of familial and idiopathic patients recruited, and small sample sizes. In addition, ance addition, ancestral background of patients in these studies, which could associate with differential<br>variant frequencies, is often not reported or accounted for. Therefore, the prevalence estimated by<br>these studies might n wariant frequencies, is often not reported or accounted for. Therefore, the prevalence estimated by<br>these studies might not be reliable.<br>In this study, we estimated the genetic prevalence (eGP) of PRKN-PD in diverse ancest variant frequencies, in the the reported of accounted the forecast of protective, these studies might not be reliable.<br>
In this study, we estimated the genetic prevalence (eGP) of PRKN-PD in diverse ancestries using<br>
heter In this study, we estimated<br>heterozygous carrier frequencies of<br>nucleotide variants and frameshift-<br>(SVs), We also estimated eGP of PR<br>ancestries and the proportions of su In this study, we extend the general of reported pathogenic PRKN variants, including small variants (single de variants and frameshift-causing small indels, referred below as SmVs) and structural variants (e also estimated heterozygous carrier and frameshift-causing small indels, referred below as SmVs) and structural variants<br>(SVs), We also estimated eGP of PRKN-PD in the USA, based on the estimates obtained in the different<br>ancestries and (SVs), We also estimated eGP of PRKN-PD in the USA, based on the estimates obtained in the different<br>ancestries and the proportions of sub-populations reported in the US census.<br> (SVS), we also estimate the proportions of sub-populations reported in the US census.<br>
An extension of PRK of PRK of PRK on the USA, based on the USA, based on the USA, based on the USA, and the USA, and the USA of PRK of ancestries and the proportions of sub-populations reported in the US census.

Compilation of reported pathogenic PRKN variants from different data sources<br>We queried pathogenic variants information from three public databases – MDSGene, ClinVar and

The pathogenic variants in the matter of the public variants in the public of the public data of the pathogene performed a systematic literature screening and curated phenotypic and genotypic data of  $PRKN$ -PD patients from OMIN (Figure 1)<br>MDSGene perform<br>PRKN-PD patient<br>(AAO) of *PRKN-P*<br>August 2022).<br>We obtained *PRI*<br>(last accessed on<br>implicated in PD<br>multiple submitt<br>guideline" (4 goli<br>We extracted *P*,<br>autosomal reces. 18–20.<br>Imed<br>18 froll<br>19 De that<br>19 star<br>18 star<br>18 star<br>18 star<br>18 star<br>18 star *PRKN-PD* patients from 174 studies<sup>11,18</sup>. We obtained the curated genotypes and ages at disease onset (AAO) of *PRKN-PD* patients from the MDSGene online database (www.mdsgene.org, last accessed on 3 August 2022).<br>We obt *PRKN-PD patients from 174 studies<sup>2222</sup>. We obtained the curated genotypes and ages at disease onset* (AAO) of *PRKN-PD patients from the MDSGene online database (www.mdsgene.org, last accessed on 3 August 2022).*<br>We obta (Aasset 2022).<br>
We obtained *PRKN* variants reported in ClinVar using the website https://www.ncbi.nlm.nih.gov/clinvar/<br>
(last accessed on 9 December 2021). We only considered pathogenic or likely pathogenic variants<br>
impl Me obtained P<br>(last accessed<br>implicated in P<br>multiple subm<br>guideline" (4 g<br>We extracted<br>autosomal rec<br>Finally, we adc<br>Finally, we adc<br>For each repol<br>and we exclud (last accessed on 9 December 2021). We only considered pathogenic or likely pathogenic variants<br>implicated in PD that were rated with at least two gold stars (including review status "criteria provided,<br>multiple submitters

(implicated in PD that were rated with at least two gold stars (including review status "criteria proventitiple submitters, no conflicts" (2 gold stars), "reviewed by expert pane!" (3 gold stars) and "praguideline" (4 gold imultiple submitters, no conflicts" (2 gold stars), "reviewed by expert panel" (3 gold stars) and "practice guideline" (4 gold stars)).<br>
We extracted *PRKN* small variants that were reported in OMIM (https://omim.org/) to guideline" (4 gold stars)).<br>We extracted *PRKN* small variants that were reported in OMIM (https://omim.org/) to be causin<br>autosomal recessive PD (last accessed on 13 December 2021).<br>Finally, we added the variant c.535-3A> We extracted *PRKN* sma<br>autosomal recessive PD (I<br>Finally, we added the vari<br>For each reported pathog<br>and we excluded those w<br>another. We further asse<br>inferred the GRCh38 genomenclature or the GRC<br>Reported *PRKN* pathog We entrance which small variants that were reported in three lapanese families <sup>21</sup>, and<br>autosomal recessive PD (last accessed on 13 December 2021).<br>Finally, we added the variant c.535-3A>G which has recently been reported Finally, we added the variant c.535-3A>G which has recently be<br>For each reported pathogenic mutation, we assessed the evidence<br>another. We further assessed the pathogenicity of these mutations<br>another. We further assessed Finally, we added the variant c.535-3A>G which has recently been reported in three Japanese families <sup>21</sup>.<br>For each reported pathogenic mutation, we assessed the evidence from the three mutation databases,<br>and we excluded Face it for filling and the second in the second second second in the secon and we excluded those with conflicting evidence – reported as pathogenic in one source while benign in<br>another. We further assessed the pathogenicity of these mutations based on ACMG guidelines<sup>22</sup>. We<br>inferred the GRCh38

Reported PRKN pathogenic variants observed in human reference databases<br>To obtain allele frequencies (AF) of the PRKN pathogenic variants in the general population, we

another. We further assessed the pathogenicity of these mutations based on ACMG guidelines<sup>22</sup>. We<br>inferred the GRCh38 genomic coordinates of these mutations, when not available, based on the HGVS<br>nomenclature or the GRCh3 another. We further assessed the pathogenicity of these mutations based on ACMG guidelines<sup>21</sup>. We<br>inferred the GRCh38 genomic coordinates of these mutations, when not available, based on the HGVS<br>nomenclature or the GRCh3 momenclature or the GRCh37 genomic coordinates.<br> **Reported PRKN pathogenic variants observed in human reference databases**<br>
To obtain allele frequencies (AF) of the *PRKN* pathogenic variants in the general population, we<br> Reported *PRKN* pathogenic variants observed i<br>To obtain allele frequencies (AF) of the *PRKN* p<br>interrogated five data sources (Figure 1).<br>We interrogated gnomAD v2.1.1 to obtain allele<br>African/African American and Japane Interrogated five data sources (Figure 1).<br>We interrogated five data sources (Figure 1).<br>We interrogated gnomAD v2.1.1 to obtain allele frequencies of SmVs in most ancestries, except<br>African/African American and Japanese a Me interrogated five data sources (Figure 1).<br>We interrogated gnomAD v2.1.1 to ob<br>African/African American and Japanese<br>Americans, we interrogated gnomAD v3<br>WGS data generated by the Tohoku Med<br>Omics Reference Panel (jMorp African/African American and Japanese ancestries. For allele frequencies of SmVs in African/African<br>Americans, we interrogated gnomAD v3. For allele frequencies of SmVs in Japanese, we interrogated<br>WGS data generated by th Americans, we interrogated gnomAD v3. For allele frequencies of SmVs in Japanese, we interrogated<br>WGS data generated by the Tohoku Medical Megabank Organization (ToMMO) via the Japanese Multi-<br>Omics Reference Panel (jMorp) Americans, we interrigated garding to the function papel and representation (ToMMO) via the Japanese Multi-Omics Reference Panel (jMorp) website<sup>23</sup>. Both the gnomAD v3 African/African American cohort and the 500 microsof Omics Reference Panel (jMorp) website $^{23}$ . Both the gnomAD v3 African/African American cohort and the  $\sim$ 

Thus providing more accurate allele frequencies in those populations.<br>GnomAD v2 SmVs: We obtained the gnomAD v2.1.1 liftover SmV data from the gnomAD website<br>(https://gnomad.broadinstitute.org/downloads, last assessed on 2 GnomAD v2 SmVs: We obtained the gnomAD v2.1.1 liftover SmV<br>(https://gnomad.broadinstitute.org/downloads, last assessed on 20 Ju<br>from 141,456 unrelated individuals from different ancestral groups (2)<br>whole genomes), lifted Ghttps://gnomad.broadinstitute.org/downloads, last assessed on 20 July 2023). This data was generated<br>from 141,456 unrelated individuals from different ancestral groups (125,748 whole exomes and 15,708<br>whole genomes), lift

(https://grow.org/downloads/inducidinals/inducidinstitute on 141,456 unrelated individuals from different ancestral groups (125,748 whole exomes and 15,708 whole genomes), lifted over from the GRCh37 to the GRCh38 human re manuscript and states individuals of African and States in different and the computer whole genomes, lifted over from the GRCh37 to the GRCh38 human reference sequence.<br>GnomAD v3 SmVs: We obtained the gnomAD v3.1.2 SmV all GnomAD v3 SmVs: We obtained the gnomAD v3.1.2 SmV allele information from v<br>sequences of 20,744 individuals of African genetic ancestry. All the genome sequences we<br>GRCh38.<br>TOMMO SmVs: We accessed allele frequencies of SmV GRCh38.<br>GRCh38.<br>TOMMO SmVs: We accessed allele frequencies of SmVs in Japanese ancestry from TOMMO<br>(https://jmorp.megabank.tohoku.ac.jp/). These summary statistics were generated from whole genome

SRCh38.<br>TOMMO SmVs: We accessed allele frequencies of SmVs in Japanese ancestry from TOMMO<br>(https://jmorp.megabank.tohoku.ac.jp/). These summary statistics were generated from whole genome<br>sequencing data of 54,000 Japanes TOMMO<br>(<u>https://j</u><br>sequencii (https://jmorp.megabank.tohoku.ac.jp/). These summary statistics were generated from whole genome<br>sequencing data of 54,000 Japanese individuals.<br>In order to incorporate structural variant population data, we leveraged two

(and the sequencing data of 54,000 Japanese individuals.<br>In order to incorporate structural variant population data, we leveraged two data sources. We first<br>obtained the SV data recently made available from 10,847 genomes sequencing data of 54,000 pm<br>In order to incorporate structural variant popu<br>obtained the SV data recently made available<br>leveraged SVs information from WGS of 8,300 コード こうしょう In the set of the SV data recently made available from 10,847 genomes in gnomAD v2.1<sup>24</sup>. We also<br>Ieveraged SVs information from WGS of 8,300 Japanese available on the TOMMO jMorp<sup>25</sup>. We linked<br>the SVs reported in gnomAD leveraged SVs information from WGS of 8,300 Japanese available on the TOMMO jMorp<sup>25</sup>. We linked<br>the SVs reported in gnomAD v2.1 and TOMMO-8.3KJPN-SV to variants found in PRKN-PD patients based<br>on the genomic region overla leveraged SVs information from WGS of 8,300 Japanese available on the TOMMO jMorp<sup>25</sup>. We linked<br>the SVs reported in gnomAD v2.1 and TOMMO-8.3KJPN-SV to variants found in PRKN-PD patients based<br>on the genomic region overla on the genomic region overlap. We assumed that the SVs were the same if they were of the same<br>variant type (deletion or duplication) and had the same consequences on the cDNA (e.g. complete<br>deletion of exon 3).<br>**Supplement** on the genomic region overlap. The assumed that the same consequences on the cDNA (e.g. complete deletion of exon 3).<br>deletion of exon 3).<br>**Supplementary Table 1** shows the number of sequenced individuals in each database. variant type (determined important) and had the same consequences on the same (e.g. complete<br>deletion of exon 3).<br>Calculation of heterozygous pathogenic variants carrier frequencies

# Calculation of heterozygous pathogenic variants carrier frequencies

Supplementary Tab<br>Calculation of het<br>We inferred the fre Supplementary Table 1 shows the number of sequenced individuals in each database.<br>
Calculation of heterozygous pathogenic variants carrier frequencies<br>
We inferred the frequencies of heterozygous pathogenic variants carrie

reference databases mentioned above.<br>For each reported pathogenic *PRKN* variant observed in gnomAD v2, gnomAD v3 or TOMMO, we calculated the carrier frequency in each population using the following formula: reference databases mentioned above.<br>For each reported pathogenic *PRKN* variant observed in gnomAD v2, gnomAD v3 or TOMMO, we<br>calculated the carrier frequency in each population using the following formula:<br>Carrier frequ For each reported pathogenic PRKN v<br>calculated the carrier frequency in each<br>Carrier frequency of variant v For each reported pathogenic PRKN variant observed in gnomAD v2, gnomAD v3 or TOMMO, we<br>calculated the carrier frequency in each population using the following formula:<br>Carrier frequency of variant v  $(CF_v) = \frac{Number of heterozygous carriers}{Number of individuals}$ 

Carrier frequency of variant v  $(CF_v) = \frac{Number\ of\ heterozygous\ carriers}{Number\ of\ individuals}$ Carrier frequency of variant v  $(CF_v) = \frac{Namber}{Number of individuals}$ wamber of marbiaaais

For each variant and each ancestry population, we calculated the 95% confidence interval (CI) of the CF<br>using the Wilson score.<br>To estimate the genetic prevalence of *PRKN*-PD in each ancestral population, we used the foll

### Estimation of genetic prevalence in various ancestries

using the Wilson score.<br> **Estimation of genetic prevalence in various ancestries**<br>
To estimate the genetic prevalence of *PRKN*-PD in each ancestral population, we used the following<br>
approach as described elsewhere<sup>8</sup>.

To estimate the genetic prevalence of PRKN-PD in each ancestral population, we used the following<br>approach as described elsewhere<sup>8</sup>.<br>Assuming all biallelic variant combinations were possible, pathogenic and fully penetran approach as described elsewhere<br>Assuming all biallelic variant comb<br>all possible combinations of two<br>prevalence. Considering each pos<br>variant, i.e. homozygote genotyp ii<br>.<br>e<br>e Assuming all possible combinations of two known pathogenic variants in the calculation of *PRKN-PD* genetic<br>prevalence. Considering each possible pair of variants i and j (where i and j can be either the same<br>variant, i.e. prevalence. Considering each possible pair of variants i and j (where i and j can be either the same<br>variant, i.e. homozygote genotype, or different variants, i.e. compound heterozygote genotypes), we<br>estimated the geneti variant, i.e. homozygote genotype, or different variants, i.e. compound heterozygote genotypes), we<br>estimated the genetic prevalence in each ancestry using the formula:

Estimated genetic prevalence = 
$$
\frac{\sum (CF(i) * CF(j))}{4}
$$

varianted the genetic prevalence in each ancestry using the formula:<br>
Estimated genetic prevalence =  $\frac{\sum (CF(i) * CF(j))}{4}$ <br>
This estimated genetic prevalence of *PRKN*-PD is equivalent to the probability of heterozygous carri Estimated genetic prevalence  $= \frac{\sum (CF(i) * CF(j))}{4}$ <br>This estimated genetic prevalence of *PRKN*-PD is equivalent to the p<br>having offspring with biallelic pathogenic variants. Estimated genetic prevalence =  $\frac{\Delta U}{\Delta t}$ <br>imated genetic prevalence of *PRKI*<br>offspring with biallelic pathogenic v:<br>d the CI lower bounds and upper k<br>th ancestry population.

This estimated genetic prevalence of PRKN-PD is equivalent to the probability of heterozygous carriers<br>having offspring with biallelic pathogenic variants.<br>We used the CI lower bounds and upper bounds of the variants' CF t We used the CI lower bounds and upper bounds of<br>CI in each ancestry population.<br>Inference of *PRKN*-PD genetic prevalence in tl CI in each ancestry population.<br>Inference of PRKN-PD genetic prevalence in the USA and world-wide<br>We inferred the PRKN-PD genetic prevalence in the USA based on 1) the estimated PRKN-PD genetic

### Inference of PRKN-PD genetic prevalence in the USA and world-wide

**COMPLATATION, POPULATION**<br> **Inference of PRKN-PD genet**<br>
We inferred the PRKN-PD gene<br>
prevalence in the different an |
|
|} prevalence in the different ancestries, and 2) the major population proportions reported in the US<br>census as well as the proportions of European and Asian sub-populations reported in the community<br>surveys.<br>The proportions preventive in the difference in the difference in the different and the difference in the different surveys.<br>The proportions of major populations in the USA were obtained from the 2021 US Census data<br>(https://www.census.go

surveys.<br>The proportions of major populations in the USA were obtained from the 2021 US Census data<br>(https://www.census.gov/quickfacts/fact/table/US/PST045221, last accessed on 20 July 2022). We also<br>obtained the proportio surveys.<br>The prop<br>obtained<br>website (https://www.census.gov/quickfacts/fact/table/US/PST045221, last accessed on 20 July 2022). We also<br>obtained the proportion of self-reported Europeans and Asians from the American Community Survey<br>website (https://www.cens (https://www.census.gov/programs-surveys/acs/data.html).<br>We matched these self-reported races and ethnicities or countries of origin to the genetic ancestries<br>defined by the human reference population databases and inferre

obtained the proportion of self-reported enterpedual information train the function entangle website (https://www.census.gov/programs-surveys/acs/data.html).<br>We matched these self-reported races and ethnicities or countrie We matched these self-reported races and ethnicities or countries<br>defined by the human reference population databases and inferred<br>the USA. We applied the same approach, using the 2022 Revisi defined by the human reference population databases and inferred the *PRKN*-PD genetic prevalence in<br>the USA. We applied the same approach, using the 2022 Revision of World Population Prospects<br>7 the USA. We applied the same approach, using the 2022 Revision of World Population Prospects  $t$  and  $t$  appears the same approach, using the 2022 Revision of Population Properties  $\tau$ 

prevalence of PRKN-PD.<br>prevalence of PRKN-PD. prevalence of PRKN-PD.

Reported pathogenic PRKN-PD variants<br>We first compiled and curated the reported human pathogenic PRKN-PD variants from three databases

WE first complied and curated the reported numan pathogente PRKN-D variants from three databases<br>(MDSGene, ClinVar, OMIM), and investigated their allele frequency in five large diverse population<br>datasets (gnomAD v2, gnomA (Matasets (gnomAD v2, gnomAD v2-SV, gnomAD v3, ToMMO-54K and ToMMO-8.3K-SV) (Figure 1).<br>A total of 159 distinct disease-causing genetic variants were identified in the *PRKN* gene, after consolidating structural variants ( datasets (gnomAD v2, gnomAD v2-SV, gnomAD v3, ToMMO-54K and ToMMO-8.3K-SV) (Figure 1).<br>A total of 159 distinct disease-causing genetic variants were identified in the *PRKN* gene<br>consolidating structural variants (SV) with A total of 159 distinct disease-causing generic variants were identified in the PRKN gene, after<br>consolidating structural variants (SV) with same consequences on coding sequence (**Supplementary**<br>Table 2). Among these, 72 ( Table 2). Among these, 72 (45.3%) were SVs (large deletions, large duplications, large indels, other large rearrangements), 48 (30.2%) were predicted loss-of-function (pLoF) single nucleotide variants (SmVs: nonsense, fram rearrangements), 48 (30.2%) were predicted loss-of-function (pLoF) single nucleotide variants (SmVs:<br>nonsense, frameshift, or splice variants), 33 (20.8%) were missense variants and 6 were other types of<br>SmVs.<br>Out of the 1

rearrangements), 48 (20.2%) were missense variants and 6 were other types of<br>smVs.<br>Out of the 159 disease-causing *PRKN* variants, 57 were present in the population datasets. This includes<br>17 SVs (12 large deletions and 5 nonsense, manusians, etc., price variants), security, its simulation and a material system.<br>SmVs.<br>Out of the 159 disease-causing *PRKN* variants, 57 were present in the population datasets. This includes<br>17 SVs (12 large d Out of<br>17 SVs<br>presen<br>We int present in gnomAD v2, gnomAD v3 or ToMMO.<br>We interrogated MDSGene<sup>18</sup>, which provides patient-level genotype and phenotype data, to investigate

the frequency of individual variants among PRKN-PD patients, identify the variants most frequently reported in patients and ensure that those variants were present in the population datasets to contribute to the genetic prevalence estimations. \<br>t<br>r We interrogated MDSGene<sup>29</sup>, which provides patient-level genotype and phenotype data, to investigate<br>the frequency of individual variants among *PRKN*-PD patients, identify the variants most frequently<br>reported in patient the frequency of individual variants among PRKN-PD patients, identify the variants most frequency<br>reported in patients and ensure that those variants were present in the population datasets to<br>contribute to the genetic pre

contribute to the genetic prevalence estimations.<br>As previously described<sup>11</sup>, the most common variant in *PRKN*-PD patients in MDSGene was<br>c.(171+1\_172-1)\_(412+1\_413-1)del, resulting in a complete deletion of exon 3. Thi As previously described<sup>11</sup>, the most commor<br>c.(171+1\_172-1)\_(412+1\_413-1)del, resulting in a<br>99 patients out of 701 (14%) index *PRKN*-PD pa<br>reported Asian ethnicity: 58.7% *PRKN*-PD index<br>contract to 22.7% Asians report As previously described<sup>21</sup>, the most common variant in *PRKN-PD* patients in MDSGene was<br>c.(171+1\_172-1)\_(412+1\_413-1)del, resulting in a complete deletion of exon 3. This SV was observed in<br>99 patients out of 701 (14%) c.(171+1\_172-1)\_(412+1\_413-1)del, resulting in a complete deletion of exon 3. This SV was observed in<br>99 patients out of 701 (14%) index *PRKN*-PD patients. This variant was more frequent in patients of<br>reported Asian ethn Patients out of 701 (14%) index PRKN-PD index patients. This variant was more requerie in patients of<br>reported Asian ethnicity: 58.7% *PRKN*-PD index patients with this variant were of Asian ethnicity, in<br>contrast to 32.7 reported Asian ethnicity: 58.7% PRKN-PD index patients with this variant were of Asian ethnicity, in<br>contrast to 32.7% Asians reported among *PRKN*-PD index patients not carrying this SV (P=7x10<sup>-4</sup>). Six<br>additional SVs le contrast to 32.7% Asians reported among *PRKN*-PD index patients not carrying this SV (P=7x10<sup>-</sup>). Six<br>additional SVs leading to complete deletion of exon 3 were reported in seven Caucasian homozygous or<br>compound heterozyg compound heterozygous *PRKN*-PD index patients, bringing the total number of index patients with an<br>exon 3 deletion to 106 (15%). In gnomAD v2.1, nine distinct SVs causing complete deletion of exon 3<br>were reported. The mos exon 3 deletion to 106 (15%). In gnomAD v2.1, nine distinct SVs causing complete deletion of exon 3 were reported. The most frequent of those nine SVs (DEL\_6\_76253) had an allele frequency (AF) ranging between 0.01% in Eur exon 3 deletion 1 deletion 10 deletion to 106 (15%). In given a member 2012 manufacture deletion to 100% in the<br>setween 0.01% in European and African/African American ancestries, and 0.08% in East Asians. When<br>combining al between 0.01% in European and African/African American ancestries, and 0.08% in East Asians. When<br>combining all nine SVs reported in gnomAD v2.1, the allele frequency of exon 3 deletions across<br>9  $\frac{1}{2}$  combining all nine SVs reported in gnomAD v2.1, the allele frequency of exon 3 deletions across. 9 combine SVs reported in gnomand in gnomand in gradient, the alleged in gradient state  $\overline{a}$ 

different populations was 0.07%, ranging between 0.04% in African/African Americans, 0.08% in<br>Europeans, 0.12% in East Asians and 0.16% in Latino/admixed American ancestries. Two SVs<br>corresponding to a complete deletion of

Europeans, 1921 in East Asian and 1920 in East, and 2007, annual American ancestrate the Tele<br>
an aggregated allele frequency of 0.04%.<br>
The second most common pathogenic variant was the frameshift variant c.155delA (p.Asn an aggregated allele frequency of 0.04%.<br>The second most common pathogenic variant was the frameshift variant c.155delA (p.Asn52Metfs\*29),<br>reported in 78 (11%) *PRKN-PD* index patients. Most patients (92.1%) carrying this The second most common pathogenic va<br>reported in 78 (11%) *PRKN*-PD index<br>Caucasians (in contrast to 40.9% Cauca:<br>P=3.6x10<sup>-12</sup>). In the general population The second most common pathogens can also an anti-mathematic variant variant wave reported in 78 (11%) *PRKN-PD* index patients. Most patients (92.1%) carrying this variant were Caucasians (in contrast to 40.9% Caucasians reported in 78 (11%) PRKN-PD index patients. Most patients (32.1%) carrying this variant were<br>Caucasians (in contrast to 40.9% Caucasians reported in the overall PRKN-PD index patients cohort,<br>P=3.6x10<sup>-12</sup>). In the genera  $P=3.6x10^{-12}$ ). In the general population (based on gnomAD v2.1.1, gnomAD v3, and JMorp), this frameshift variant was most frequent among Latino or admixed Americans (AF=0.11%) followed by non-<br>Finnish Europeans (AF=0.02 frameshift variant was most frequent among Latino or admixed Americans (AF=0.11%) followed by non-<br>Finnish Europeans (AF=0.024%) and African/African Americans (AF=0.012%). This variant was absent in<br>Asian reference populat Finnish Europeans (AF=0.024%) and African/African Americans (AF=0.012%). This variant was absent in

# $\mathsf{F}$ stimoshift variant was most frequent among Latino or admixed  $\mathsf{F}$  and  $\mathsf{F}$  followed by non-latino  $\mathsf{F}$ Estimated genetic prevalence of PRKN-PD in the general population

Asian reference population control (AF=0, in ground and ToMMO).<br>Estimated genetic prevalence of *PRKN*-PD in the general popu<br>We next leveraged the allele frequencies of the 57 *PRKN* known dise We next leveraged the allele requencies of the 57 PRAN known disease-causing Shivs and Svs present<br>in the population datasets to estimate the genetic prevalence (eGP) of *PRKN*-PD in four major<br>genetically-defined populati in the population datasets to estimate the genetic prevalence (eGP) or PRKN-PD in four major<br>genetically-defined populations in gnomAD, and in Japanese using the ToMMO data (Figure 2 and<br>Supplementary Table 3). We also cal Supplementary Table 3). We also calculated the genetic prevalence in six non-Finnish European (NFE) subpopulations, in two East Asian genetically-defined subpopulations, in Ashkenazi Jews, and in South Asians represented i subpopulations, in two East Asian genetically-defined subpopulations, in Ashkenazi Jews, and in South<br>Asians represented in gnomAD v2 (Supplementary Table 3). However, we note that some<br>(sub)populations eGP are based only Substitute of the Substitute of the Substitute of the Substitute of the Substitutions of the Subpopulations eGP are based only on SmVs and do not include SVs information due to absence of SV<br>The subpopulations eGP are base

Asians represented in gnomAD v2 (Supplementary Table 3). However, we note that some<br>(sub)populations eGP are based only on SmVs and do not include SVs information due to absence of SV<br>reference data in those populations.<br>A (sub)populations established only on SMVs and the based of the theorem on SMVs and the based on SPS cl=4-165 per<br>As shown on Figure 2, East Asians had the highest eGP (24 in 100,000 individuals, 95% Cl=4-165 per<br>100,000 in reference data in those populations.<br>
As shown on Figure 2, East Asians had the highest eGP (24 in 100,000 individuals, 95% Cl=4-165 per<br>
100,000 individuals), largely driven by SVs, particularly deletions of exon 2, exon

As shown on Figure 2, East Asians had the highest eGP (24 in 100,000 individuals, 33% CI=4 103 per 100,000 individuals), largely driven by SVs, particularly deletions of exon 2, exon 3 and exons 2-3 (Figure 3).<br>
The secon 100,000 individuals), largely driven by Svs, particularly deletions of exon 2, exon 3 and exons 2-3 (Figure 3).<br>
100,000 individuals). We note that NFE had the highest lower bound eGP (11 in 100,000 NFE individuals vs 4 in رد<br>The<br>ind<br>200<br>the Individuals). We note that NFE had the highest lower bound eGP (11 in 100,000 NFE individuals vs 4 in 100,000 East Asians). The three most frequent variants contributing to eGP in NFE included deletions of exon2, c.823C>T individuals). The three most frequent variants contributing to eGP in NFE included deletions of<br>exon2, c.823C>T (p.Arg275Trp) and c.1310C>T (p.Pro437Leu) (**Figure 3**). The two SmVs were also among<br>the top five most frequen exon2, c.823C>T (p.Arg275Trp) and c.1310C>T (p.Pro437Leu) (**Figure 3**). The two SmVs were also among<br>the top five most frequent variants contributing to eGP in all other non-Asian populations (**Figure 3**).<br>c.823C>T and c.1 exon2, c.823C>T (p.Arg275Trp) and c.1310C>T (p.T0437Led) (Figure 3). The two Shivs were also among<br>the top five most frequent variants contributing to eGP in all other non-Asian populations (Figure 3).<br>c.823C>T and c.1310C  $\text{c.823C>T}$  and  $\text{c.1310C>T}$  were reported in 68 (9.7%) and 13 (1.9%) index patients in MDSGene,  $\text{c.823C>T}$  and  $\text{c.1310C>T}$  were reported in 68 (9.7%) and 13 (1.9%) index patients in MDSGene, c.823C-T and c.1310C-T were reported in 10<br>The contract in the following index patients in MDSGene, patients in MDSGene, patients in MDSGene, patients in<br>The contract in MDSGene, patients in MDSGene, patients in MDSGene, p

respectively, and predominantly in Caucasians (85.7% and 100% of index patients with those pathogenic<br>variants and ethnicity were Caucasians, respectively).<br>The lowest eGP was observed in Japanese, with eGP=6 per 1 million 3 per 100,000 individuals) and was largely driven by SVs as well as two SmVs (c.535-3A>G and c.110C>T/<br>p.Pro37Leu). (Figure 2, Figure 3 and Supplementary Table 3). Because our Japanese estimates are<br>derived from a large co The lowest eGP was observed in Japanese, with eGP=6 per 1 million individuals (95% CI = 1 per 1 million-3 p. Pro37Leu). (Figure 2, Figure 3 and Supplementary Table 3). Because our Japanese estimates are<br>derived from a large cohort of individuals (thus enhancing the accuracy of the carrier frequencies), and<br>because of the sma p. Frostedy. (Figure 2, Figure 3 and Supplementary Table 3). Because our Japanese estimates are<br>derived from a large cohort of individuals (thus enhancing the accuracy of the carrier frequencies), and<br>because of the small derived from a large consideration of the small number of individuals of Japanese ancestry within the East Asian cohort in gnomAD<br>v2 (N=76 Japanese out 9,977 East Asians in the gnomADv2 SmV dataset; unknown number of Japan because of the small number of individuals of the small number of such that send to the grounds of 1,208 East Asians in gnomADv2 SV dataset), the high eGP in East Asians is likely driven by high eGP in non-Japanese East As out of 1,208 East Asians in gnomADv2 SV dataset), the high eGP in East Asians is likely driven by high<br>eGP in non-Japanese East Asians.<br>We next determined the genetic prevalence of PRKN-PD in the USA by extrapolating from

out of 1,208 Eq. 1,208 Me<br>1,208 We next determined the genetic prevalence of *PRKN*-PD in the USA by extrapolating from the reported endian<br>We next determined the genetic<br>proportions of races and ethnicitie<br>estimated genetic prevalence of proportions of races and ethnicities in the US Census and American Community Survey, coupled with the estimated genetic prevalence of the corresponding genetically-defined populations described above.<br>Based on this analysi proportional defined genetic prevalence of the corresponding genetically-defined populations described above.<br>
Based on this analysis, we estimated the genetic prevalence of *PRKN*-PD in the USA to be 18 per 100,000<br>
indiv

Based on this analysis, we estimated the genetic prevalence of *PRKN*-PD in the USA to be 18 per 100,000<br>individuals (95% CI: 7-68 per 100,000 individuals) (**Figure 2** and **Supplementary Table 3**).<br>Using the same approach Based on this analysis, we estimated the genetic prevalence of PRKN-PD in the USA to be 18 per 100,000<br>individuals (95% Cl: 7-68 per 100,000 individuals) (Figure 2 and Supplementary Table 3).<br>Using the same approach and th Il sing the same approach and the proportion of races and ethnicities in the 2022 Repoulation Prospects published by the United Nations, we estimated the world-wide ger of *PRKN*-PD to be 13 per 100,000 individuals (95% CI Population Prospects published by the United Nations, we estimated the world-wide genetic prevalence<br>of *PRKN*-PD to be 13 per 100,000 individuals (95% Cl: 3-70 per 100,000 individuals) (**Figure 2** and<br>**Supplementary Table** of *PRKN*-PD to be 13 per 100,000 individuals (95% Cl: 3-70 per 100,000 individuals) (**Figure 2** and **Supplementary Table 3**). Supplementary Table 3).<br>Supplementary Table 3). Supplementary Table 3).

In the literature, at least 1,530 symptomatic *PRKN*-PD patients have been reported to date<sup>26</sup>. Based on<br>previously published studies, Wasner et al suggested that *PRKN*-PD represents around 0.5–1% of PD<br>patients across patients across all AAO groups<sup>13</sup>. Based on the data from the Global Burden of Disease Study, Dorsey et al estimated that around 6.2 million individuals worldwide have PD, representing 0.082% of the total population (7.6 patients across all AAO groups<sup>13</sup>. Based on the data from the Global Burden of Disease Study, Dorsey et<br>al estimated that around 6.2 million individuals worldwide have PD, representing 0.082% of the total<br>population (7.6 population (7.6 billion)<sup>14</sup>. Assuming *PRKN* biallelic pathogenic variants contribute to 0.5–1% of PD<br>patients across all age groups, the prevalence of *PRKN*-PD would be 4-8 in 1 million individuals<br>worldwide.<br>*PRKN*-PD, patients across all age groups, the prevalence of *PRKN*-PD would be 4-8 in 1 million individuals<br>worldwide.<br>*PRKN*-PD, like many other rare diseases, is better understood in specific human populations<br>where the majority o

patients across an age groups, the prevalence of PRKN-PD would be 4-8 in 1 million individuals<br>worldwide.<br>PRKN-PD, like many other rare diseases, is better understood in specific human populations<br>where the majority of stu *PRK*<br>where the r<br>of Asian (39<br>understand PRKN-PD, like many other rare diseases, is better diderstood in specific human populations<br>he majority of studies have been conducted. In MDSGene, the majority of *PRKN*-PD patients are<br>it (39%; primarily from Iran, China of Asian (39%; primarily from Iran, China or Japan) or Caucasian (31%) ancestry<sup>11</sup>. In order to better understand the genetic epidemiology of *PRKN*-PD in diverse populations, we have investigated the *PRKN*-PD genetic pr understand the genetic epidemiology of *PRKN*-PD in diverse populations, we have investigated the *PRKN*-PD genetic prevalence eGP within eight genetically-defined ancestries and eight genetically-defined subpopulations by PRKN-PD genetic prevalence eGP within eight genetically-defined ancestries and eight genetically-<br>defined subpopulations by leveraging large-scale genetic datasets from five human reference population<br>databases.<br>When estim PRKN-PD genetic prevalence eGP within eight genetically-defined ancestries and eight genetically-<br>defined subpopulations by leveraging large-scale genetic datasets from five human reference population<br>databases.<br>When estim

databases.<br>
When estimating the genetic prevalence of a disease, it is imperative to first meticulously curate the<br>
reported disease-causing variants from confirmed patients, carefully investigate the frequency of the<br>
kno When estir<br>reported d<br>known path !<br>ל<br>ו reported disease-causing variants from confirmed patients, carefully investigate the frequency of the<br>known pathogenic variants and variants types in reported patients, and take into account the ancestries<br>represented with known pathogenic variants and variants types in reported patients, and take into account the ancestries<br>represented within the cohort of reported patients, in order to inform the eGP analysis design and<br>interpretation of r represented within the cohort of reported patients, in order to inform the eGP analysis design and<br>interpretation of results. Previous GP studies in rare diseases have largely focused on diseases<br>predominantly caused by pr interpretation of results. Previous GP studies in rare diseases have largely focused on diseases<br>predominantly caused by protein-coding SmVs and relied on gnomAD as the only source of SmV carrier<br>frequencies. Our study des interpredominantly caused by protein-coding SmVs and relied on gnomAD as the only source of SmV carrier<br>frequencies. Our study design differs from previous GP studies on two fronts. First, in order to enhance<br>the accuracy prequencies. Our study design differs from previous GP studies on two fronts. First, in order to enhance<br>the accuracy of the SmV carrier frequencies, we harnessed data from three reference databases<br>(gnomAD v2, gnomAD v3 a frequency of the SmV carrier frequencies, we harnessed data from three reference databases<br>(gnomAD v2, gnomAD v3 and jMorp), and selected for each variant and each ancestry, the carrier<br>frequency from the reference databas the accuracy of the SmV carrier frequencies, we harnessed data from three reference databases (gnomAD v2, gnomAD v3 and jMorp), and selected for each variant and each ancestry, the carrier frequency from the reference data (gnomabled the utilization of carrier frequency data from 54,000 Japanese, an ancestry that is significantly<br>underrepresented in gnomAD. Secondly, as described by MDSGene and other publications, SVs<br>contribute significantl enabled the utilization of carrier frequency data from 54,000 Japanese, an ancestry that is significantly<br>underrepresented in gnomAD. Secondly, as described by MDSGene and other publications, SVs<br>contribute significantly t enabled the unifiable the utilization of carrier frequency data from 54,000 Lippanese, and other publications, SVs<br>contribute significantly to the genetic cause of familial *PRKN*-PD <sup>11,21</sup>. To address this observation, w contribute significantly to the genetic cause of familial *PRKN*-PD <sup>11,21</sup>. To address this observation, we<br>probed two reference datasets – gnomAD v2-SV and ToMMO 8.3K SV - that provide carrier frequency of<br>SVs, and high probed two reference datasets – gnomAD v2-SV and ToMMO 8.3K SV - that provide carrier frequency of<br>SVs, and highlight the role of SVs in the genetic prevalence of *PRKN*-PD. problems the reference datasets – gnoming to 2-S and ToMMO 8.3K and highlight the role of SVs in the genetic prevalence of *PRKN*-PD.<br>12 Svs, and highlight the role of Svs in the genetic prevalence of PRKN-PD.

 $\begin{array}{c} \n\frac{1}{2} & \frac{1}{2} \\ \n\frac{1}{2} & \frac{1}{$ Three key findings can be derived from our results. First, our results suggest that the prevalence of  $PRKN$ -PD may be substantially higher than previously thought, with an eGP of 13 per 100,000 individuals worldwide (95% C results indicate that PRKN-PD eGP varies significantly between ancestries ranging from only 6 in 1 worldwide (95% CI: 3-70 per 100,000 individuals) (Figure 2 and Supplementary Table 3). Second, our results indicate that *PRKN*-PD eGP varies significantly between ancestries ranging from only 6 in 1 million Japanese (95% results indicate that PRKN-PD eGP varies significantly between ancestries ranging from only 6 in 1<br>million Japanese (95% Cl = 1 per 1 million-3 per 100,000 individuals) to 22 in 100,000 NFE (95% Cl=11-64<br>per 100,000 indivi mer 100,000 individuals) and 24 in 100,000 non-Japanese East Asians (95% CI=4-165 per 100,000<br>
individuals). Third, our analysis supports that the variants types contributing to the disease etiology<br>
varies greatly between individuals). Third, our analysis supports that the variants types contributing to the disease etiology<br>varies greatly between ancestries, with the highest frequency of disease-causing SVs in non-Japanese<br>East Asians and h varies greatly between ancestries, with the highest frequency of disease-causing SVs in non-Japanese<br>East Asians and highest frequency of disease-causing SmVs in NFE (Figure 2 and Figure 3). Interestingly,<br>while the *PRKN* East Asians and highest frequency of disease-causing SmVs in NFE (Figure2 and Figure 3). Interestingly, while the *PRKN*-PD literature includes a significant proportion of patients from Japan<sup>11,21</sup>, our estimates suggest East Asians and mightest riequency or disease-causing SmVs in ME (Figure2 and Figure 3). Interestingly, while the *PRKN*-PD literature includes a significant proportion of patients from Japan<sup>11,21</sup>, our estimates suggest while the PRKN-PD literature includes a significant proportion of patients from Japan<sup>22</sup>, our estimates<br>suggest that the genetic prevalence of the disease is low in the Japanese population. The majority of the<br>Japanese in suggest that the general presentation in the disease is preparately preparation. The disparement in the suggest is a low in the TOMMO 8.3KJPN-SV dataset and contributed to our genetic prevalence estimate with an aggregate particularly deletions of exon 3, exon 4 and exons 3-4 (50/86 alleles, 58%). Those three deletions were<br>present in the TOMMO 8.3KJPN-SV dataset and contributed to our genetic prevalence estimate with an<br>aggregate heterozyg present in the TOMMO 8.3KJPN-SV dataset and contributed to our genetic prevalence estimate with an aggregate heterozygous carrier frequency of 0.18% (Figure 3). A recent study reporting pathogenic variants in a large Japan present in the TOMMO scarrier frequency of 0.18% (Figure 3). A recent study reporting pathogenic<br>variants in a large Japanese familial PD cohort reported the pathogenic SmV c.535-3A>G in multiple<br>families<sup>21</sup>. This splice variants in a large Japanese familial PD cohort reported the pathogenic SmV c.535-3A>G in multiple<br>families<sup>21</sup>. This splice variant was the most frequent pathogenic SmV in the TOMMO 54K dataset and<br>contributed to our gene families<sup>21</sup>. This splice variant was the most frequent pathogenic SmV in the TOMMO 54K dataset and<br>contributed to our genetic prevalence estimate with a carrier frequency of 0.28% (Figure 3). Our findings<br>suggest that th contributed to our genetic prevalence estimate with a carrier frequency of 0.28% (Figure 3). Our findings suggest that the high proportion of Japanese patients described in the literature may be due to ascertainment bias a suggest that the high proportion of Japanese patients described in the literature may be due to<br>ascertainment bias and may not reflect the percentage of Japanese patients in the global patient<br>population. Even larger, high suggest assertainment bias and may not reflect the percentage of Japanese patients in the global patient<br>population. Even larger, high-quality, reference SV frequency data in Japanese will be required to<br>validate this find as a population. Even larger, high-quality, reference SV frequency data in Japanese will be required to<br>validate this finding.<br>It is important to make the distinction between the genetic prevalence, which reflects the prop

populate this finding.<br>
population. Even in the sense of the genetic prevalence, which reflects the proportion of<br>
individuals that carry two pathogenic variants, and the disease prevalence, which reflects the proportion<br> It is important to ma<br>individuals that carry<br>of symptomatic patio individuals that carry two pathogenic variants, and the disease prevalence, which reflects the proportion of symptomatic patients. If pathogenic variants are not fully penetrant (i.e. not all individuals carrying a biallel individuals that carrying a biallelic genotype with the variant will develop symptoms), the genetic prevalence and the disease prevalence, may be inconsistent. In the case of *PRKN*-PD, no robust information about variant biallelic genotype with the variant will develop symptoms), the genetic prevalence and the disease<br>prevalence may be inconsistent. In the case of *PRKN*-PD, no robust information about variant<br>penetrance is known. A recent biallelic general general development in the variant will develop symptoms), the general prevalence and the interaction<br>prevalence is known. A recent study in UK Biobank reported that, while carriers of two *PRKN* pathogen penetrance inay be inconsistent. In the case of PRKN-PD, no robust information about variant<br>penetrance is known. A recent study in UK Biobank reported that, while carriers of two *PRKN* pathogenic<br>variants were more likel variants were more likely to have a Parkinson's disease diagnosis, 22 (91.7%) biallelic carriers had no symptoms<sup>27</sup>. This observation can represent incomplete ascertainment of disease status, incomplete symptoms<sup>27</sup>. This observation can represent incomplete ascertainment of disease status, incomplete<br>13<br>

13

information due to difference between current age of the individuals and age of disease onset, and/or<br>can reflect incomplete penetrance of (some of) the variants. Noteworthily, deletions of exons 2, 3 and 4<br>have been repor have been reported with a frequency higher than expected in the general population for pathogenic<br>variants. In the gnomAD v2 dataset (N=10,847 samples), SVs leading to complete deletions of exon 2,<br>exon 3 or exon 4 (which variants. In the gnomAD v2 dataset (N=10,847 samples), SVs leading to complete deletions of exon 2,<br>exon 3 or exon 4 (which were reported in 57, 99 and 67 homozygous and compound heterozygous index<br>patients in MDSGene) ha exon 3 or exon 4 (which were reported in 57, 99 and 67 homozygous and compound heterozygous index<br>patients in MDSGene) have an aggregate carrier frequency of 0.8%, 0.15% and 0.06% respectively in<br>gnomAD v2 (Figure 3). Mor patients in MDSGene) have an aggregate carrier frequency of 0.8%, 0.15% and 0.06% respectively in<br>gnomAD v2 (Figure 3). More work will be required to better assess SVs in *PRKN* (and gene-level SV<br>status per individual) an gnomAD v2 (Figure 3). More work will be required to better assess SVs in PRKN (and gene-level SV<br>status per individual) and their consequences, and to better understand the penetrance of PRKN<br>pathogenic variants, both SmVs gnomab v2 (rigure 3). More work will be required to better assess SVs in PRKN (and gene-level SV<br>status per individual) and their consequences, and to better understand the penetrance of *PRKN*<br>pathogenic variants, both Sm pathogenic variants, both SmVs and SVs. For Instance, deletion of exons 3-4 was frequently reported as<br>homozygous pathogenic variant in AR *PRKN*-PD patients in MDSGene. However, some studies have<br>suggested that deletions homozygous pathogenic variant in AR *PRKN*-PD patients in MDSGene. However, some studies have suggested that deletions of both exons 3 and 4 maintain the reading frame and are associated with a later onset, milder disease, suggested that deletions of both exons 3 and 4 maintain the reading frame and are associated with a<br>later onset, milder disease, in contrast to deletion of exon 3 that leads to a frameshift with more severe<br>consequences<sup>2</sup> later onset, milder disease, in contrast to deletion of exon 3 that leads to a frameshift with more severe

Several limitations stem from the current state of knowledge regarding the genetic etiology of the<br>disease. The understanding of pathogenic variants in the PRKN gene may be less comprehensive in consequences<sup>28,29</sup>.<br>Several limitations<br>disease. The unde<br>certain populations c/<br>(<br>( disease. The understanding of pathogenic variants in the *PRKN* gene may be less comprehensive in certain populations that are under-represented in studies, potentially leading to an underestimation of genetic prevalence i disease. The understanding of pathogenic variants in the PRKN gene may be less complemensive in<br>certain populations that are under-represented in studies, potentially leading to an underestimation of<br>genetic prevalence in genetic prevalence in those populations. It is important to note that our analysis was limited to reported<br>PRKN pathogenic variants observed in early-onset recessive forms of PD patients. Variants predicted to<br>be pathogeni PRKN pathogenic variants observed in early-onset recessive forms of PD patients. Variants predicted to<br>be pathogenic using in silico tools and not (yet) reported in patients were not included in our analysis.<br>While adding be pathogenic using *in silico* tools and not (yet) reported in patients were not included in our analysis.<br>While adding these variants would further increase the eGP estimates, caution is warranted as it may result in inf be pathogenic using in sinco tools and not (yet) reported in patients were not included in our analysis.<br>While adding these variants would further increase the eGP estimates, caution is warranted as it may<br>result in inflat The state variants would further the settlems of the continuation of false positive pathogenic variants as discussed above.<br>Other limitations relate to the current knowledge about genetic variation in the general human<br>pop Other limitations relate to the current knowledge about genetic variation in the general human<br>population and the different ancestries. While massive efforts, like gnomAD, have drastically improved<br>our understanding of the population and the different ancestries. While massive efforts, like gnomAD, have drastically improved<br>our understanding of the spectrum of genetic variants on the human genome and their frequencies is<br>various ancestries, pour understanding of the spectrum of genetic variants on the human genome and their frequencies is<br>various ancestries, and have remarkably benefited genetic diagnosis and other aspects of human health,<br>ongoing and future various ancestries, and have remarkably benefited genetic diagnosis and other aspects of human health,<br>ongoing and future efforts to further improve the cataloguing of human genetic variation in diverse<br>populations will be ongoing and future efforts to further improve the cataloguing of human genetic variation in diverse<br>populations will be required to further improve the accuracy of genetic prevalence estimations. Human<br>reference population populations will be required to further improve the accuracy of genetic prevalence estimations. Human reference population databases contain genotypic data generated using different sequencing platforms (whole exome sequen population databases contain genotypic data generated using different sequencing platforms<br>(whole exome sequencing vs whole genomes) and the number of samples from each ancestry are<br>different. This may result in variabilit (whole exome sequencing vs whole genomes) and the number of samples from each ancestry are different. This may result in variability in the accuracy of carrier frequency in different populations and, as a consequence, resu different. This may result in variability in the accuracy of carrier frequency in different populations and, as a consequence, result in variability in the accuracy of eGP between populations. In our analysis, we as a consequence, result in variability in the accuracy of eGP between populations. In our analysis, we fix

14

of individuals sequenced to account for this limitation.<br>In addition, limited information about SVs is currently available. This is particularly relevant to *PRKN*-PD<br>in which SVs contribute significantly to the disease et In addition, limited information about SVs is currently<br>in which SVs contribute significantly to the disease e<br>data sources that capture information about SVs, and<br>and from jMorp into our analysis, future increase In which SVs contribute significantly to the disease etiology. While we carefully considered available data sources that capture information about SVs, and incorporated SVs information from gnomAD v2.1 and from jMorp into in which SVs contribute significantly to the disease energy. Thus its station, contribute a antitative<br>data sources that capture information about SVs, and incorporated SVs information from gnomAD v2.1<br>and from jMorp into and from jMorp into our analysis, future increases in sample sizes in diverse populations and<br>improvements in technologies to call SVs and assess their consequences on the protein's function will be<br>required.<br>It is crucial

improvements in technologies to call SVs and assess their consequences on the protein's function will be<br>required.<br>It is crucial to also emphasize that the methodology employed in this study was specifically tailored to<br>es improvements in tequired.<br>It is crucial to also emphasize that the methodology employed in this study was specifically tailored to<br>estimate the genetic prevalence of autosomal recessive PRKN-PD caused by biallelic variants required.<br>It is crucia<br>estimate<br>Complicat<br>more proi It is crucial to also emphasize that the methodology employed in the clasy the specifically tailors in *PRKN*.<br>Complicating the analysis is the fact that *PRKN* is one of the largest gene in the genome, rendering it<br>more p Complicating the analysis is the fact that *PRKN* is one of the largest gene in the genome, rendering it<br>more prone to genetic variation. Future studies will be imperative to enhance our understanding of the<br>functional con Complicating the analysis is the fact that PRKN is one of the largest gene in the genome, rendering it<br>more prone to genetic variation. Future studies will be imperative to enhance our understanding of the<br>functional conse more prone to genetic variation. Future studies will be imperative to enhance our understanding of the functional consequences of variants in *PRKN* and their role in AR PRKN-PD.<br>In summary, the methodology used in this st

functional consequences of variants in PRKN and their role in AR PRKN-PD.<br>In summary, the methodology used in this study has proven effective in int<br>the prevalence of multiple rare recessive diseases. As demonstrated the<br>P  $\frac{1}{t}$ In summary, the methodology necessive diseases. As demonstrated through our investigation into *PRKN*-PD, it serves as a powerful tool to complement epidemiology research, offering valuable insights for patient identificat the prevalence of multiple rare recessive and substrated the assumence and  $S$ . The anticologistic into prevalence of prevalence of prevalence of prevalence of prevalence in the prevalence of prevalence of prevalence of pr Frakty-PD, it serves as a powerful tool to complement epidemiology research, offering valuable insights for patient identification strategies and informing the design of clinical trials. for patient identification strategies and informing the design of clinical trials.

### Acknowledgements

**NA** 

### Data availability

na<br>Dat<br>All<br>sup  $\begin{array}{c} \n\frac{1}{2} & \frac{1}{2} \\ \n\frac{1}{2} & \frac{1}{$ 

### Author contributions

All data generated or analyzed during this study are included in this published article and its<br>supplementary information files.<br>DD and EHMW: study design, execution, analysis, writing, and editing of the final version of supplementary information files.<br> **Author contributions**<br>
DD and EHMW: study design, execution, analysis, writing, and editing of the final version of the<br>
manuscript. |
|
| r manuscript.<br>CD, and WQ: analysis and review of the final version of the manuscript<br>SC, TH, and JP: review and editing of the final version of the manuscript.<br>DP, JL and GP: study design, review and editing of the final ver

 $CD$ , and WQ: analysis and review of the final version of the manuscript<br>SC, TH, and JP: review and editing of the final version of the manuscript.

CC, TH, and JP: review and editing of the final version of the manuscript<br>DP, JL and GP: study design, review and editing of the final version of the<br>**Funding**<br>Takede Davelement Conter Americas, lee provided funding

SC, TH, and GP: study design, review and editing of the final version of the manuscript.<br>DP, JL and GP: study design, review and editing of the final version of the<br>**Funding**<br>Takeda Development Center Americas, Inc. provid

### Funding

Punding<br>Depending<br>Takeda Development Center Americas, Inc. provided funding.<br>Compating interacts.  $\begin{array}{c} \n\frac{1}{2} \quad \text{if} \quad \frac{1}{2} \quad \text{if} \quad \frac{$ 

### Competing interests

Takeda Development Center Americas, Competing interests<br>Takeda Development Center Americas, Inc. provided funding.<br>The employees of Takeda Pharmaceuticals Company Lim (<br>|
| DD, The proposes of Takeda Pharmaceuticals Company Limited. At time of the study, EHMW was an employee of Takeda Development Center Americas, Inc.; DD, EHMW, CD, JL, SC, TH, JP, DP and GP are stockholders of Takeda Pharmac employee of Takeda Development Center Americas, Inc. DD, EHMW, CD, JL, SC, TH, JP, DP and GP<br>are stockholders of Takeda Pharmaceuticals Company Limited. All authors declare no non-financial<br>competing interests. are stockholders of Takeda Pharmaceuticals Company Limited. All authors declare no non-financial are stockholders of Takeda Pharmaceuticals Company Limited. All authors declare no non-financial<br>competing interests.<br>. competing interests.

1. Haendel, M. et al. How many rare diseases are there? Nat Rev Drug Discov 19, 77–78 (2020).<br>
2. Melnikova, I. Rare diseases and orphan drugs. Nat Rev Drug Discov 11, 267–268 (2012).<br>
3. Nguengang Wakap, S. et al. Estimat 2. Melnikova, I. Rare diseases and orphan drugs. Nut Rev Drug Discov 11, 207–200 (2012).<br>
3. Mguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: and<br>
the Orphanet database. *Eur J Hum Gene* the Orphanet database. Eur J Hum Genet 28, 105–173 (2020).<br>4. Chunn, L. M. *et al.* Estimation of ENPP1 deficiency gen<br>literature review and population databases. *Orphanet J Rare D*<br>5. Gao, J., Brackley, S. & Mann, J. P.

3. Nguengang Wakap, 3. et al. Estimating cumulative point prevalence of rare diseases. analysis of<br>the Orphanet database. *Eur J Hum Genet* 28, 165–173 (2020).<br>4. Chunn, L. M. *et al.* Estimation of ENPP1 deficiency geneti 4. Chunn, E. W. et al. Estimation of ENTT1 deficiency genetic prevalence using a comprenensive<br>literature review and population databases. Orphanet J Rare Dis 17, 421 (2022).<br>5. Gao, J., Brackley, S. & Mann, J. P. The glob If eractic review and population databases. Orphanet J Rare Dis 17, 421 (2022).<br>
5. Gao, J., Brackley, S. & Mann, J. P. The global prevalence of Wilson diseas<br>
sequencing data. *Genet Med* **21**, 1155–1163 (2019).<br>
7. Park,

Sequencing data. Genet Med 21, 1155–1163 (2015).<br>6. Park, K. S. Carrier frequency and predicted g<br>general population database. Mol Genet Metab Rep<br>7. Zhao, T., Fan, S. & Sun, L. The global carrier f<br>Schulman syndrome. BMC

general population database. Mor denet Metab Rep 27, 100734 (2021).<br>7. Zhao, T., Fan, S. & Sun, L. The global carrier frequency and genet<br>Schulman syndrome. *BMC Genom Data* 22, 50 (2021).<br>8. Hanany, M., Rivolta, C. & Shar

sequencing data. *Genet Med* **21**, 1155–1163 (2019).<br>
6. Park, K. S. Carrier frequency and predicted genetic prevalence of Pompe disease based on a<br>
general population database. *Mol Genet Metab Rep* **27**, 100734 (2021).<br> general population database. *Mol Genet Metab Rep* 27, 100734 (2021).<br>7. Zhao, T., Fan, S. & Sun, L. The global carrier frequency and genetic prevalence of Upshaw-<br>Schulman syndrome. *BMC Genom Data* 22, 50 (2021).<br>8. Hana

protein degradation and rupture of the outer mitochondrial membrane. J Biol Chem 286, 19630-19640 Schuman syndrome. *BMC* Genom Data 22, 50 (2021).<br>8. Hanany, M., Rivolta, C. & Sharon, D. Worldwid<br>autosomal recessive inherited retinal diseases. *Proc No*<br>9. Kitada, T. *et al.* Mutations in the parkin gene co<br>*Nature* 8. Hanany, M., Rivolta, C. & Sharon, D. Worldwide carrier frequency and genetic prevalence of autosomal recessive innerited retinal diseases. Proc Ndir Actu Scr 0 3 A 117, 2710–2710 (2020).<br>
9. Kitada, T. *et al.* Mutations in the parkin gene cause autosomal recessive juvenile parkins<br> *Nature* **392**, 605–608 (1998 9. Kitada, T. et al. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism.<br>
10. Yoshii, S. R., Kishi, C., Ishihara, N. & Mizushima, N. Parkin mediates proteasome-dependent<br>
protein degradation and r Nature 332, 605–608 (1998).<br>10. Moshii, S. R., Kishi, C.,<br>protein degradation and rupt<br>(2011).<br>11. Kasten, M. *et al.* Gene<br>PINK1, DJ1: MDSGene System<br>12. Doherty, K. M. *et al.* F<br>13. Wasner, K., Grünewa protein degradation and rupture of the outer mitochondrial membrane. *J Biol Chem* 286, 19630–19<br>(2011).<br>11. Kasten, M. *et al.* Genotype-Phenotype Relations for the Parkinson's Disease Genes Parkin,<br>PINK1, DJ1: MDSGene Sy

(2011).<br>
11. Kasten, M. *et al.* Genotype-Phenotype Relations for the Parkinson's Disease Genes P<br>
PINK1, DJ1: MDSGene Systematic Review. *Mov Disord* **33**, 730–741 (2018).<br>
12. Doherty, K. M. *et al.* Parkin disease: a cl

protein degradation and rupture of the outer infloctionial membrane. J Biol Chem 286, 19690–19640<br>
11. Kasten, M. *et al.* Genotype-Phenotype Relations for the Parkinson's Disease Genes Parkin,<br>
PINK1, DJ1: MDSGene Systema 11. Kasten, M. et al. Genotype-Phenotype Relations for the Parkinson's Disease Genes Parkin,<br>PINK1, DJ1: MDSGene Systematic Review. *Mov Disord* 33, 730–741 (2018).<br>12. Doherty, K. M. *et al.* Parkin disease: a clinicopath PINK1, DJ1: MD3Gene Systematic Review. Mov Disord 33, 730–741 (2016).<br>12. Doherty, K. M. *et al.* Parkin disease: a clinicopathologic entity? JAM<br>13. Wasner, K., Grünewald, A. & Klein, C. Parkin-linked Parkinson's dise<br>pat 13. Bonerty, K. M. et al. Farkin disease. a clinicopathologic entity? JAMA Neurol 70, 571–579 (2015).<br>13. Wasner, K., Grünewald, A. & Klein, C. Parkin-linked Parkinson's disease: From clinical insights to<br>pathogenic mechan

pathogenic mechanisms and novel therapeutic approaches. Neurosci Res 155, 34–35 (2020).<br>14. Dorsey, E. R. & Bloem, B. R. The Parkinson Pandemic-A Call to Action. JAMA Neurol 7.<br>15. Lücking, C. B. *et al.* Association betwe

pathogenic mechanisms and novel therapeutic approaches. *Neurosci Res* 159, 34–39 (2020).<br>14. Dorsey, E. R. & Bloem, B. R. The Parkinson Pandemic-A Call to Action. *JAMA Neurol* 75, 9–10<br>(2018).<br>15. Lücking, C. B. *et al.* 14. Borsey, E. R. & Bloem, B. R. The Parkinson Pandemic-A call to Action. JAMA Neurol 73, 9–10<br>(2018). Lücking, C. B. et al. Association between early-onset Parkinson's disease and mutations in the<br>parkin gene. N Engl J Me

(2012).<br>15.<br>parkin g<br>16.<br>Parkinso<br>17.<br>for gene<br>1116–1 15. Lucking, C. B. et al. Association between early-onset Parkinson's disease and mutations in the<br>parkin gene. N Engl J Med 342, 1560–1567 (2000).<br>16. Tan, M. M. X. et al. Genetic analysis of Mendelian mutations in a larg parkin gene. *N Engl 3 Med 342*, 1560–1567 (2000).<br>16. Tan, M. M. X. *et al.* Genetic analysis of Mer<br>Parkinson's disease study. *Brain* **142**, 2828–2844 (2<br>17. Alcalay, R. N. *et al.* Frequency of known mi<br>for genetic cou 16. Tran, M. M. A. et al. Genetic analysis of Mendelian mutations in a large OK population-based<br>Parkinson's disease study. *Brain* 142, 2828–2844 (2019).<br>17. Alcalay, R. N. *et al.* Frequency of known mutations in early-o Parkinson's disease study. Brain 142, 2828–2844 (2019).<br>17. Alcalay, R. N. *et al.* Frequency of known mutatior<br>for genetic counseling: the consortium on risk for early of<br>1116–1122 (2010).<br>18. Lill, C. M. *et al.* Launchi 17. Alcalay, N. W. et al. Frequency of known mutations in early onset Parkinson disease. Impileation<br>for genetic counseling: the consortium on risk for early onset Parkinson disease study. Arch Neurol 67,<br>1116–1122 (2010).

for genetic counseling: the consortium on risk for early onset Farkinson disease study. Arch Neurol 67,<br>1116–1122 (2010).<br>18. Lill, C. M. *et al.* Launching the movement disorders society genetic mutation database<br>(MDSGene 18. Lill, C. M. *et*<br>(MDSGene). *Mov D.*<br>19. Landrum, N<br>*Nucleic Acids Res* 44<br>20. Hamosh, A.<br>Inheritance in Man<br>**33**, D514-517 (2005 18. Lin, C. M. et al. Launching the movement disorders society genetic mutation database<br>(MDSGene). *Mov Disord* **31**, 607–609 (2016).<br>19. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically re (MDSGene). Mov Disord 31, 607–609 (2010).<br>19. Landrum, M. J. *et al.* ClinVar: public a<br>*Nucleic Acids Res* 44, D862-868 (2016).<br>20. Hamosh, A., Scott, A. F., Amberger, J.<br>Inheritance in Man (OMIM), a knowledgebase<br>33, D5 19. Landrum, M. J. et al. Chilvar: public archive of interpretations of enheally relevant variants.<br>
20. Landron, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian<br>
Inheritance in Man (O Nucleic Acids Res 44, D862-868 (2010).<br>20. Hamosh, A., Scott, A. F., Amber<br>Inheritance in Man (OMIM), a knowled<sub>{</sub><br>33, D514-517 (2005).<br>21. Yoshino, H. *et al.* Genotype-phe<br>Neurobiol Aging **114**, 117–128 (2022).<br>22. Richa

Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids*<br>33, D514-517 (2005).<br>21. Yoshino, H. *et al.* Genotype-phenotype correlation of Parkinson's disease with PRKN variant<br>*Neurob* Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*<br> **33**, D514-517 (2005).<br>
21. Yoshino, H. *et al.* Genotype-phenotype correlation of Parkinson's disease with PRKN variant 33, D314-517 (2005).<br>21. Yoshino, H. *et*<br>*Neurobiol Aging* **114**,<br>22. Richards, S. *et*<br>consensus recommen<br>Association for Molec 21. Yoshino, H. et al. Genotype-phenotype-correlation of Parkinson's disease with PRKN variants.<br>Neurobiol Aging 114, 117–128 (2022).<br>22. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence va Neurobiol Aging 114, 117–128 (2022).<br>22. Richards, S. *et al.* Standards an<br>consensus recommendation of the Am<br>Association for Molecular Pathology. G 22. Richards, 3. et al. Standards and guidelines for the interpretation of sequence variants: a joint<br>consensus recommendation of the American College of Medical Genetics and Genomics and the<br>Association for Molecular Path consensus recommendation of the American College of Medical College of Medical Genetics and College of Association for Molecular Pathology. *Genet Med* 17, 405–424 (2015). Association for Molecular Pathology. Genet Med 17, 405–424 (2015).

23. Tadaka, S. et al. jwist p updates in 2020. Targe emancement of multi-ormes data resources on<br>the general Japanese population. *Nucleic Acids Res* 49, D536–D544 (2021).<br>24. Collins, R. L. *et al.* A structural variation the general Japanese population. *Nucleic Acids Res* 49, D536–D544 (2021).<br>24. Collins, R. L. *et al.* A structural variation reference for medical and p<br>**581**, 444–451 (2020).<br>25. Tadaka, S. *et al.* jMorp: Japanese Multi 24. Collins, R. L. et al. A structural variation reference for medicar and population genetics. Nature<br>**581**, 444–451 (2020).<br>25. Tadaka, S. *et al.* jMorp: Japanese Multi-Omics Reference Panel update report 2023. Nucleic **381, 444–451 (2020).**<br> **25.** Tadaka, S. *et c*<br> *Res* gkad978 (2023) de<br>
26. Vollstedt, E.-J.<br>
disease worldwide. Ar<br>
27. Zhu, W. *et al.*<br>
Parkinson's disease. B<br>
28. Haylett, W. *et* 

25. Tradaka, S. et al. jwierp. Japanese Multi-Omics Reference Panel update report 2023. Multiet Acids<br>
Res gkad978 (2023) doi:10.1093/nar/gkad978.<br>
26. Vollstedt, E.-J., Kasten, M. & Klein, C. Using global team science to Res gkad978 (2023) doi:10.1093/nar/gkad978.<br>26. Vollstedt, E.-J., Kasten, M. & Klein, C. U<br>disease worldwide. Ann Neurol 86, 153–157 (20<br>27. Zhu, W. *et al.* Heterozygous PRKN muta<br>Parkinson's disease. *Brain* 145, 2077–20 disease worldwide. Ann Neurol 86, 153–157 (2019).<br>27. Zhu, W. *et al.* Heterozygous PRKN mutations are common but do not increase the risk of<br>Parkinson's disease. *Brain* 145, 2077–2091 (2022).<br>28. Haylett, W. *et al.* Alt disease worldwide. Ann Neurol 86, 153–157 (2015).<br>27. Zhu, W. *et al.* Heterozygous PRKN mutations<br>Parkinson's disease. *Brain* **145**, 2077–2091 (2022).<br>28. Haylett, W. *et al.* Altered Mitochondrial Resp<br>Function in Parki 27. Zhu, W. et al. Heterozygous PRKN mutations are common but do not increase the risk of<br>Parkinson's disease. *Brain* 145, 2077–2091 (2022).<br>28. Haylett, W. *et al.* Altered Mitochondrial Respiration and Other Features of Parkinson's disease. *Brain* 145, 2077–2091 (2022).<br>28. Haylett, W. *et al.* Altered Mitochondrial Re<br>Function in Parkin-Mutant Fibroblasts from Parkins<br>(2016).<br>29. Asai, H., Hirano, M., Kiriyama, T., Ikeda, M.<br>restoration

28. Haylett, W. et al. Altered Mitochondrial Respiration and Other Features of Mitochondrial Function in Parkin-Mutant Fibroblasts from Parkinson's Disease Patients. *Parkinsons Dis* 2016, 18:<br>(2016).<br>29. Asai, H., Hirano, Function in Parkin-Mutant Fibroblasts from Parkinson's Disease Patients. Purkinsons Dis 2016, 1819209<br>(2016).<br>29. Asai, H., Hirano, M., Kiriyama, T., Ikeda, M. & Ueno, S. Naturally- and experimentally-designed<br>restorations (2016).<br>29. Asai, H., Hirano, M., Kiriyama, T., Ikeda, M. & Ueno, S. Naturally- and experimentally-designed<br>restorations of the Parkin gene deficit in autosomal recessive juvenile parkinsonism. *Biochem Biophys*<br>Res Commun restorations of the Parkin gene deficit in autosomal recessive juvenile parkinsonism. Biochem Biophys restorations of the Parkin gene deficit in autosomal recessive juvenile parkinsonism. Biochem Biophys<br>Res Commun 391, 800–805 (2010).  $Res$  Commun 391, 800–805 (2010).

## Figures legends:

Figure 1: Study design overview. We identified 159 known pathogenic variants in PRKN implicated in<br>autosomal recessive PRKN-Parkinson disease (PRKN-PD), out of which 57 were observed in the refere<br>population databases gnom autosomal recessive PRKN-Parkinson disease (PRKN-PD), out of which 57 were observed in the reference<br>population databases gnomAD or jMorp and were included in our genetic prevalence estimations.<br>SmVs, single nucleotide var

population databases graditions and the SmVs, single nucleotide variants and frameshift-causing small indels ; SVs, structural variants.<br> **Figure 2**: Genetic prevalence estimates in various genetic ancestries, in the USA a Figure 2: Genetic prevalence estimates in various genetic ancestries, in the USA and world-wide<br>axis indicates the estimated genetic prevalence eGP (and 95% confidence interval, Wilson sco<br>100,000 individuals (Log2 scale). Figure 2: Genetic prevalence estimates in various genetic ancestries, in the OSA and world-wide. The x-<br>axis indicates the estimated genetic prevalence eGP (and 95% confidence interval, Wilson score) per<br>100,000 individual and individuals (Log2 scale). For each major genetic ancestry, two sets of eGP are shown: eGP ba<br>on SmVs (SmV), eGP based on SmVs and SVs (SmV+SV). World-wide eGP and eGP in the USA (SmV ar<br>SmV+SV) were calculated based on 100,000 individuals (Log2 scale). For each major genetic ancestry, two sets of eGP are shown: eGP based<br>on SmVs (SmV), eGP based on SmVs and SVs (SmV+SV). World-wide eGP and eGP in the USA (SmV and<br>SmV+SV) were calculated SmV+SV) were calculated based on the estimates in each genetic ancestry and the proportion of each<br>ancestry in the UN and US census. AFR, African/African American ; AMR, Latino/Mixed American ; ASJ,<br>Ashkenazi Jewish ; EAS, Smyther calculated based on the estimated in each general ances in the inepreparted of each<br>ancestry in the UN and US census. AFR, African/African American ; AMR, Latino/Mixed American ; ASJ,<br>Ashkenazi Jewish ; EAS, East A ancestry in the UN and US centerstriking interaction principles in the UN and Ashkenazi Jewish ; EAS, East Asian ; EAS-JPN: Japanese ; FIN, Finnish ; NFE, Non-Finnish European ; SAS, South Asian. SmV, single nucleotide var

South Asian. SmV, single nucleotide variants and frameshift-causing small indels; SV, structural variant.<br>
Figure 3: Heterozygous carrier frequency of pathogenic *PRKN* single nucleotide variants (SmVs) and<br>
structural var Figure 3: Heterozygous carrier frequency of pathogenic PRKN single nucleotide variants (SmVs) and<br>structural variants (SVs) in African/African Americans (A), Latino/admixed Americans (B), Non-Japanese<br>East Asians (C), Japa Figure 3: Heterozygous carrier frequency of pathogenic PRKN single nucleotide variants (ShTVS) and<br>structural variants (SVs) in African/African Americans (A), Latino/admixed Americans (B), Non-Japan<br>East Asians (C), Japane East Asians (C), Japanese (D) and Non-Finnish Europeans (E).<br>East Asians (C), Japanese (D) and Non-Finnish Europeans (E). East Asians (C), Japanese (D) and Non-Finnish Europeans (E).

### Figures

Figure 1







### Figure 3

